ICR Project Summary

Project Name	SIG	Project Summary	Name
caArray	**Microarray Repositories**	NCICB is developing their next generation microarray data repository, termed caArray. Phase I of the system, to be released in September, will be caBIG-compliant and will contain the following key features: MIAME 1.1 compliant; support for MAGE-ML import and export; utilities for the submission and retrieval of Affymetrix and GenePix native file formats; use controlled vocabularies; accessiblity through a MAGE-OM API. Many Centers in the ICR Workspace are interested in adopting caArray.	Arnie Miles Michael Showe Walter Mankowitz Wen Hwai Horng Jomol Mathew Judith Goldberg
Cancer Molecular Pages	**Genome Annotation**	This system will be based on technology originally developed by the Joint Center for Structural Genomics, to aid cancer researchers in with keeping up with all of the information being generated on a gene or gene set of interest. It is a fully functional database and automated annotation system combining automated computer-based annotations and automated data collection from experimental stations and set of web based visualization tools. This system will be ported to a caBIG-compliant architecture this year and will consume of data from caArray.	Kutbuddin Doctor Edith Zang
caWorkbench 2.0	**Data Analysis and Statistical Tools**	caWorkbench 2.0 (also known as BioWorks) is a bioinformatics platform that makes available tools for data management, analysis and visualization to the community in a convenient fashion. Some of the most fully developed capabilites of the platform include microarray data analysis, pathway analysis and reverse engineering, sequence analysis, transcription factor binding site analysis and patttern discovery. caWorkbench allows structured communication between disparate software modules which may be written with little or no knowledge of each other. For caBIG, caWorkbench 2.0 will be fully documented and integrated with caArray.	Andrea Califano
Distance-Weighted Discrimination	**Data Analysis and Statistical Tools**	DWD is a tool that performs statistical corrections to reduce systematic biases that are due to different sources of RNA, different batches of microarrays and especially different microarray platforms. This tool, currently implemented in MatLab, will be ported to an open source, caBIG-compliant platform this year.	Steve Marron Everett Zhou Michael Nebozhyn Michael Showe
FunctionExpress	**Genome Annotation**	FunctionExpress is an environment for the integrated analysis and visualization of complementary data sets. The system provides a mechanism for regular updates of integrated annotation data that can be readily associated with probes on microarrays. The application allows for plotting of raw and transformed microarray data, for viewing orthologous probesets from other experiments, and for creating literature-based gene networks. This system, currently implemented in a 2-tier, proprietary archicture will be ported to an open source, caBIG-compliant architecture this year.	Rakesh Nagarajan John Rux Harold Riethman
GenePattern	**Data Analysis and Statistical Tools**	GenePattern is a flexible analysis platform developed to support multidisciplinary genomic research programs. Its architecture and environment are expressly designed to allow rapid prototyping and integration of new technologies. For caBIG, an adapter to GenePattern will created to allow the application to read data directly from caBIO/caArray.	Ted Liefeld Michael Reich Jomol Mathew Judith Goldberg
Reactome (GKB) Data	**Pathways**	The Reactome database (formerly known as GKB), developed at Cold Spring Harbor, is a curated database of fundamental biological pathways in human which uses strict rules of assertion and evidence tracking to ensure a consistent high quality product. For caBIG, this data will be made available through a pathways data exchange format that is defined by caBIG.	Lincoln Stein Brian Gilman Alex Lash
GOMiner	**Genome Annotation**	GOMiner leverages the Gene Ontology (GO) to identify the biological processes, functions and components represented in gene lists. Instead of analyzing microarray results with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories. For caBIG, GoMiner will be adapted to a caBIG-compliant architecture and made available as a web service.	David Kane John Rux Harold Riethman
HapMap, PromoterDB	**Genome Annotation**	The HapMap database is a repository of human SNPs, their genotypes, and the linkage disequilibrium relationships among them. The Vertebrate Promoter Database (VPD) is a curated	Lincoln Stein Alex Lash (PromoterDB)

		resource for vertebrate transcription factor binding sites and their corresponding regulatory regions. Both of these datasets will be made available through caBIO in this caBIG project.	Harold Riethman (HapMap)
Magellan	**Data Analysis and Statistical Tools**	Magellan is a web-based system that allows biologists to perform complex analyses on heterogeneous data in an environment that does not require a background in computer programming or statistics. Stored data and annotations are treated as abstract entities such that arbitrary, user-defined types of information can be stored. In the context of caBIG, Magellan will consume of data from caBIO/caArray and adapt its architecture to Silver-level compliance.	David Fenstermacher Craig Street Chris Kingsley Ajay Jain
NCI-60 Data Sharing	**Microarray Repositories**	The NCI-60 are 60 diverse human cancer cell lines used by the NCI Developmental Therapeutics Program to screen >100,000 chemical compounds since 1990 for anticancer activity. The Weinstein and other laboratories have assessed potential molecular targets and modulators of activity in those cells one or a few at a time and have developed a number of "omics" databases for this data. For this project, these datasets will be made available through standards and interfaces developed/approved by caBIG.	David Kane *TBD*
Pathways Tool Development	**Pathways**	Sloan has developed resources aimed at aiding life science researchers in visualizaing and interacting with information in the context of biological pathways. These resources are BioPAX, a common exchange format for pathways data; cPath, a database focused on protein-protein interactions, and Cytoscape, a bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other stat data. Sloan will leverage these existing resources to create a set of tools that will allow users to view pathway data in the context of caBIG annotation and expression data.	Shannon McWeeney Gary Bader
PIR	**Genome Annotation**	The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research. PIR maintains the Protein Sequence Database (PSD), an annotated protein database containing over 283,000 sequences convering the entire taxonomic range. PIR is also a member of the UniProt consortium, the central international resource of protein sequence and function that unifies the PIR, Swiss-Prot, and TrEMBL databases. For caBIG, the PIR database will be grid-enabled to demonstrate how such a datasource can be discovered and consumed in a grid environment.	Cathy Wu Craig Street David Fenstermacher
Proteomics LIMS	**Proteomics**	This project will be focused on the creation of a caBIG-compliant proteomics Laboratory Information Management System (LIMS). The initial version will track the lab processes relevant to 2D gel electrophoresis but the schema will support the addition of new data types as they emerge. The availability of an open source proteomics LIMS will ultimately allow a distributed development model in which additional modules can be contributed by other centers.	Michael Ochs Tom Moloshok Steve Eschrich
Q5	**Proteomics**	The Q5 algorithm supports probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. The algorithm employs Principal Components Analysis (PCA) followed by Linear Discriminant Analysis (LDA) on whole spectrum Surface-Enhanced Laser Desorption/Ionization Time of Flight (SELDI-TOF) Mass Spectrometry (MS) data. Currently implemented in MatLab, Q5 will be ported to a caBIG-compliant, open-source architecture.	David Jewell Edith Zang Shannon McWeeney
QPACA	**Pathways**	QPACA stands for Quantitative Pathway Analysis in Cancer. It is a pathway modeling and analysis system that supports exploration of quantitative biological data in the context of a pathway description. At the center of the system is a pathway representation that enables visualization and computational analysis of pathway structure. For caBIG, QPACA will be made interoperable with Magellan, a web-based system that allows the biologist to perform complex analysis on heterogeneous data and that is being made interoperable with caBIO and caArray. QPACA will also be made interoperable with the pathway exchange standard that is being defined by caBIG.	Shannon McWeeney Ajay Jain

ICR Project Summary

RProteomics	**Proteomics**	In this project, R libraries will be developed to post-process MALDI-TOF and SELDI-TOF data. These libraries will be incorporated into a caBIG-compliant, user-friendly system that will aid cancer researchers in the processing of their TOF data. RProteomics will be developed as a grid reference implementation to demonstrate workflow in a grid enviroment.	Patrick McConnell Shannon McWeeney Craig Street David Fenstermacher
SEED	**Genome Annotation**	SEED is a framework that supports peer-to-peer annotation of genomes. Investigators can work independently on their only instances of the SEED database and synchronize their work when desired or update code versions quickly via the network. Several hundred microbial organisms are in SEED now and pipelines for high-throughput processing, e.g., BLASTing, exist for rapidly including new organisms. For caBIG, access to curated SEED data will be made available through a caBIG-compliant interface. In addition, eukaroyotic data will be added to SEED.	Edith Zang Cathy Wu Terry Disz
TrAPSS	**Translational**	TrAPSS is a system comprised of several tools that aid scientists who are searching for the genetic mutation or mutations that cause a defect or disease. The system offers support for almost all areas of a mutation discovery project from the creation and prioritization of a large candidate gene list, to the selection, ordering, and managing of primer pairs, and even support for SSCP assay results. TraAPSS will be adapted to conform to the Silver level of caBIG compliance for this year.	Ed Frank John Rux Harold Riethman Edith Zang
VISDA	**Data Analysis and Statistical Tools**	VIsual and Statistical Data Analyzer (VISDA) has been developed with the goal of revealing all of the interesting patterns in a dataset. The main application of VISDA is for multivariate cluster modeling, discovery, and visualization, particularly for data sets living in high dimensional space. Currently implemented in MatLab, VISDA will be ported to a caBIG-compliant, open-source architecture.	Joseph Wang Malik Yousef Michael Nebozhyn Michael Showe
Zebrafish microarray data sharing	**Microarray Repositories**	The Thomas Jefferson University Zebrafish Microarray Service provides microarray data generation services for the community of scientists using zebrafish as a model organism for the study of cancer. They have developed a custom microarray using the commercial Zebrafish oligo library (Compugen/Sigma-Genosys). Through this caBIG project, Thomas Jefferson will make datasets from this repository available to the research community by identifying sharable datasets, creating a web application to select these datasets, and making these datasets available via a caBIG-defined exchange format.	Alex Lash Jack London

Role	Center
Adopter	Georgetown
Adopter	Wistar
Adopter	Wistar
Adopter	Wistar
Adopter	New York
Adopter	New York
Developer	Burnham
Adopter	IFCP
Developer	Columbia
Developer	Lineberger
Developer	Lineberger
Adopter	Wistar
Adopter	Wistar
Developer	Wash U
Adopter	Wistar
Adopter	Wistar
Developer	MIT/Broad
Developer	MIT/Broad
Adopter	New York
Adopter	New York
Developer	Cold Spring
Developer	Cold Spring/Panther Informatics
Adopter	Sloan
Developer	NCI - CCR
Adopter	Wistar
Adopter	Wistar
Developer	Cold Spring
Adopter	Sloan

Adopter	Wistar
Adopter	Penn
Adopter	Penn
Developer	UC San Francisco
Developer	UC San Francisco
Developer	CCR
Adopter	TBD
Adopter	Oregon Health
Developer	Sloan
Developer	Georgetown
Adopter	Penn
Adopter	Penn
Developer	Fox Chase
Developer	Fox Chase
Adopter	Moffitt
Developer	Dartmouth
Adopter	IFCP
Adopter	Oregon Health
Adopter	Oregon Health
Developer	UC San Francisco

Developer	Duke
Adopter	Oregon Health
Adopter	Penn
Adopter	Penn
Adopter	IFCP
Adopter	Georgetown
Developer	U of C
Developer	Holden
Adopter	Wistar
Adopter	Wistar
Adopter	IFCP
Developer	Georgetown
Adopter	Wistar
Adopter	Wistar
Adopter	Wistar
Adopter	Sloan
Developer	Thomas Jefferson